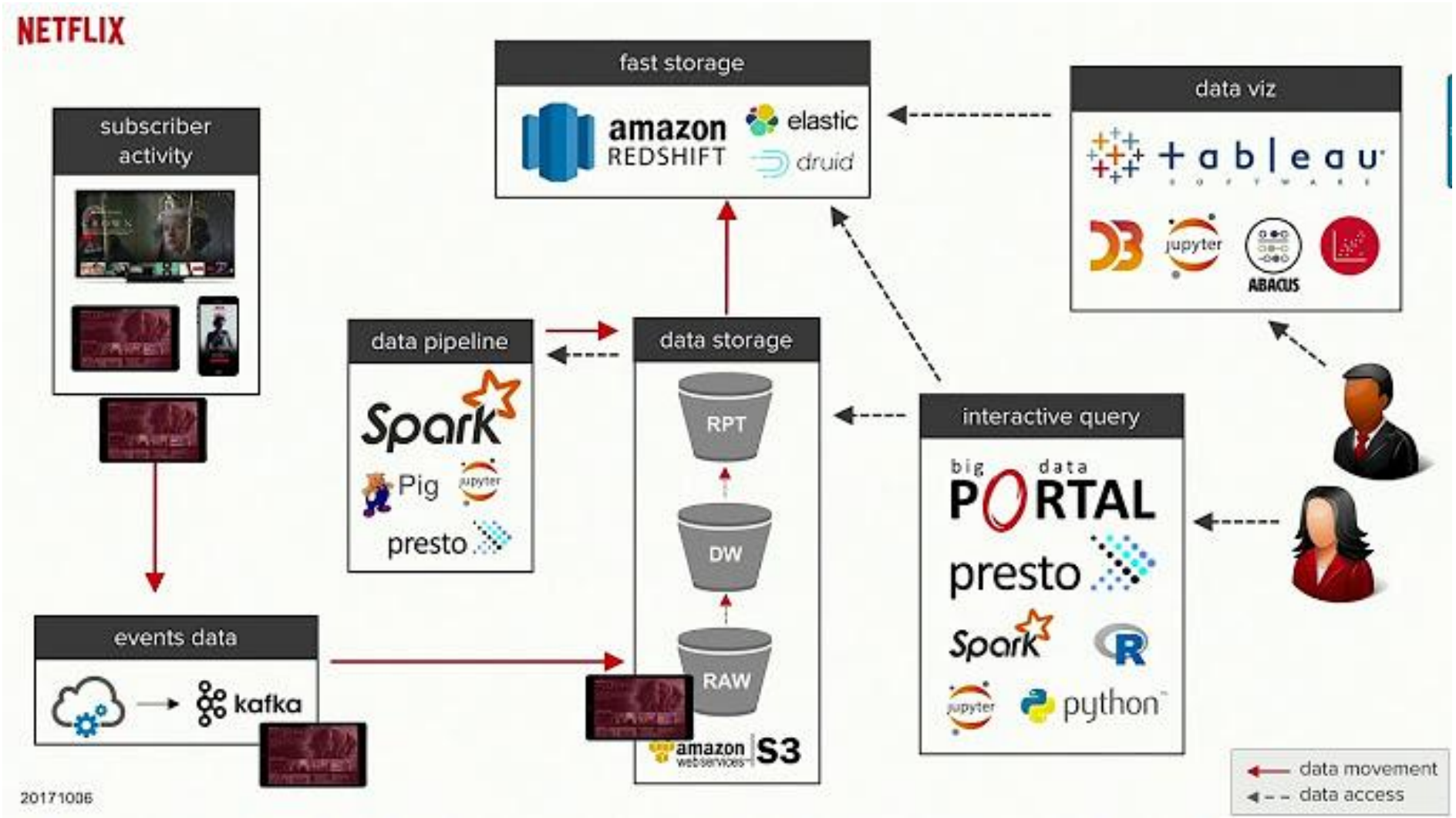
IMEN891K/IDSC501 Data Engineering

1. **Overview & Logistics**

Fall 2025

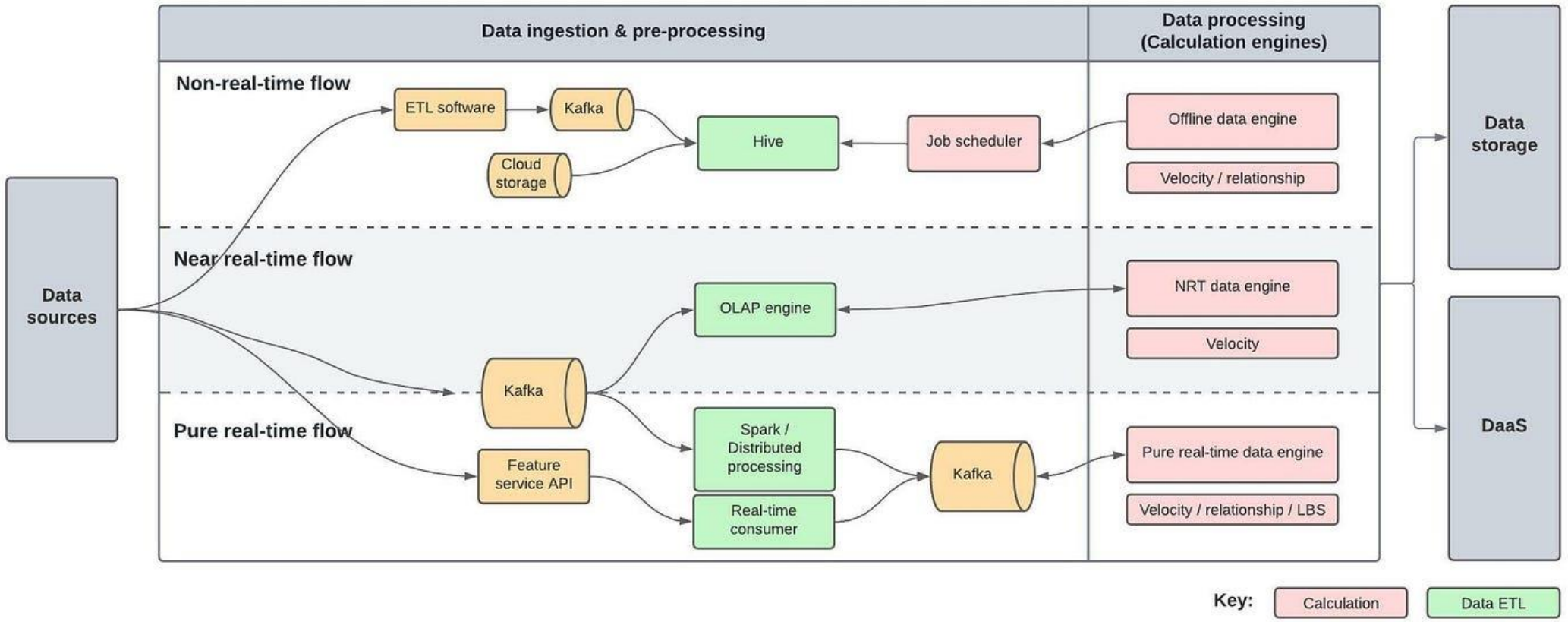
Instructor: **Sangwoo Mo**

What is Data Engineering?

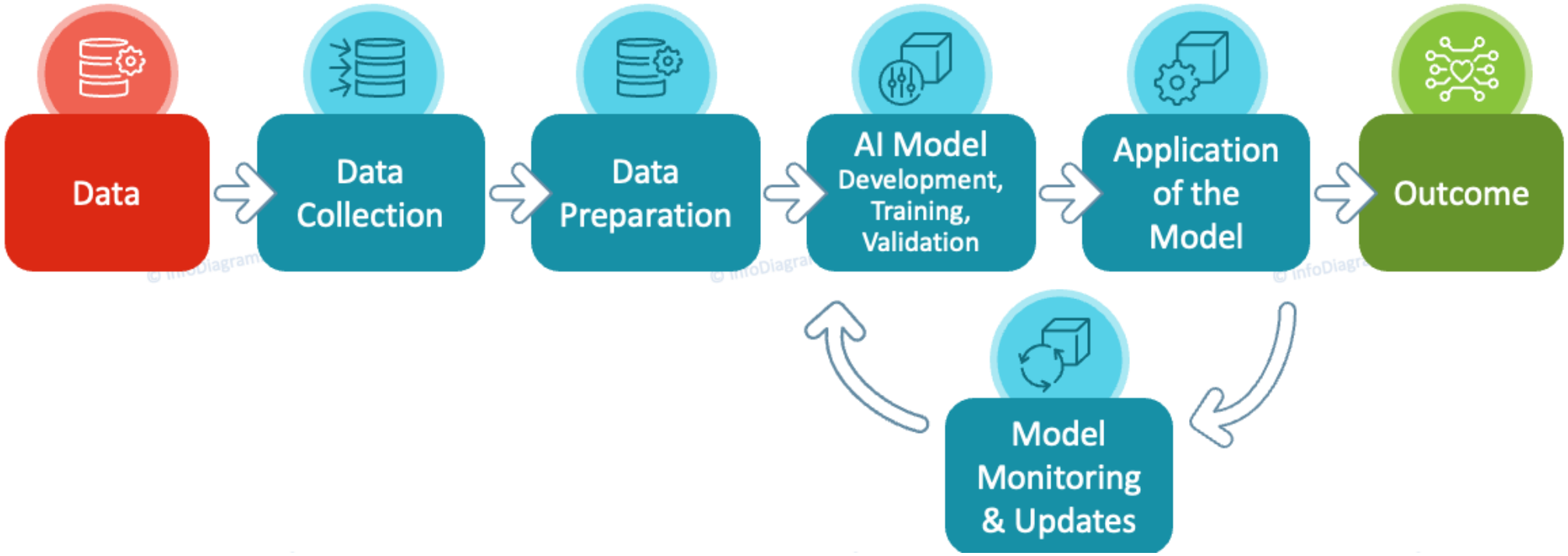


What is Data Engineering?

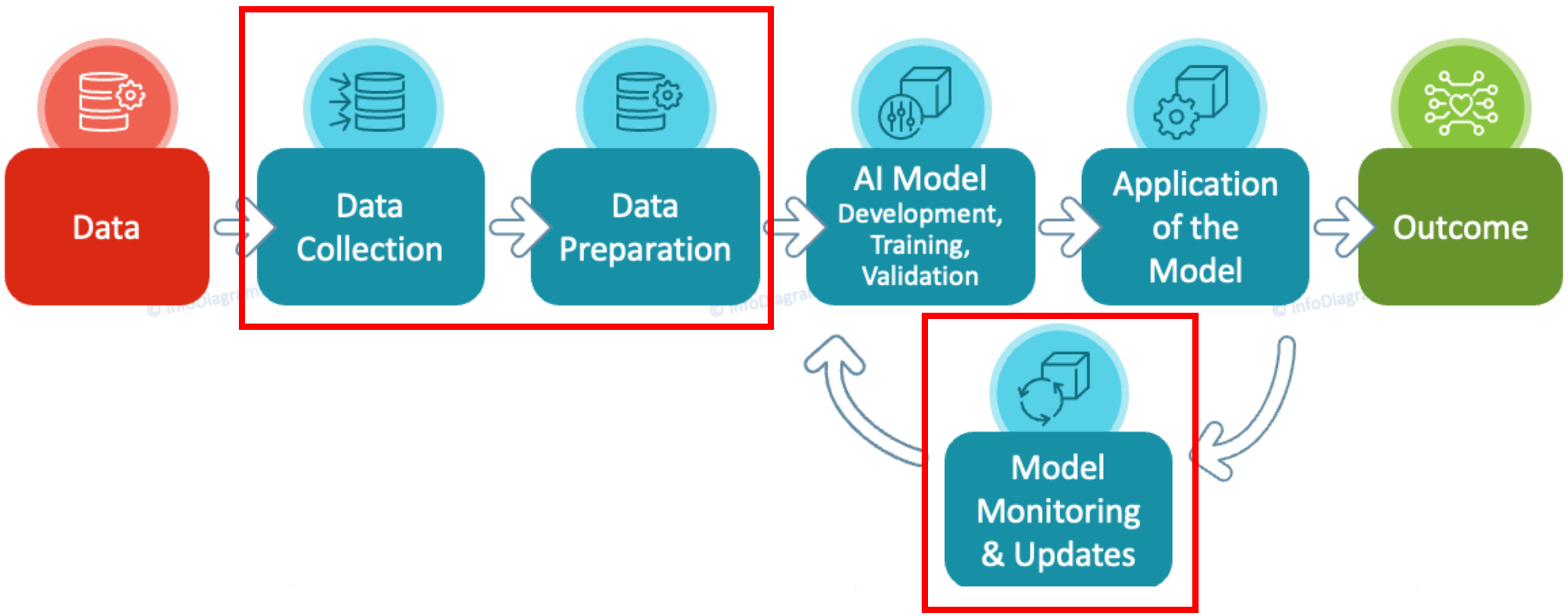
coupang eats



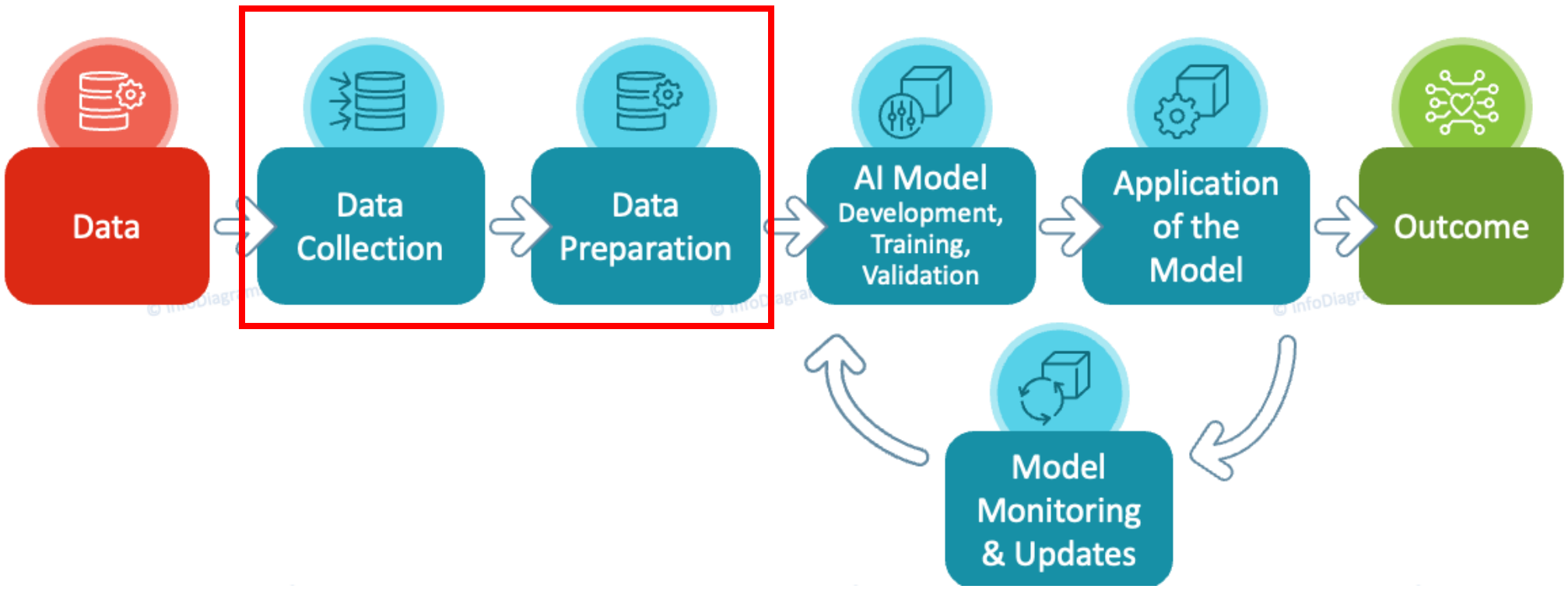
What is Data Engineering?



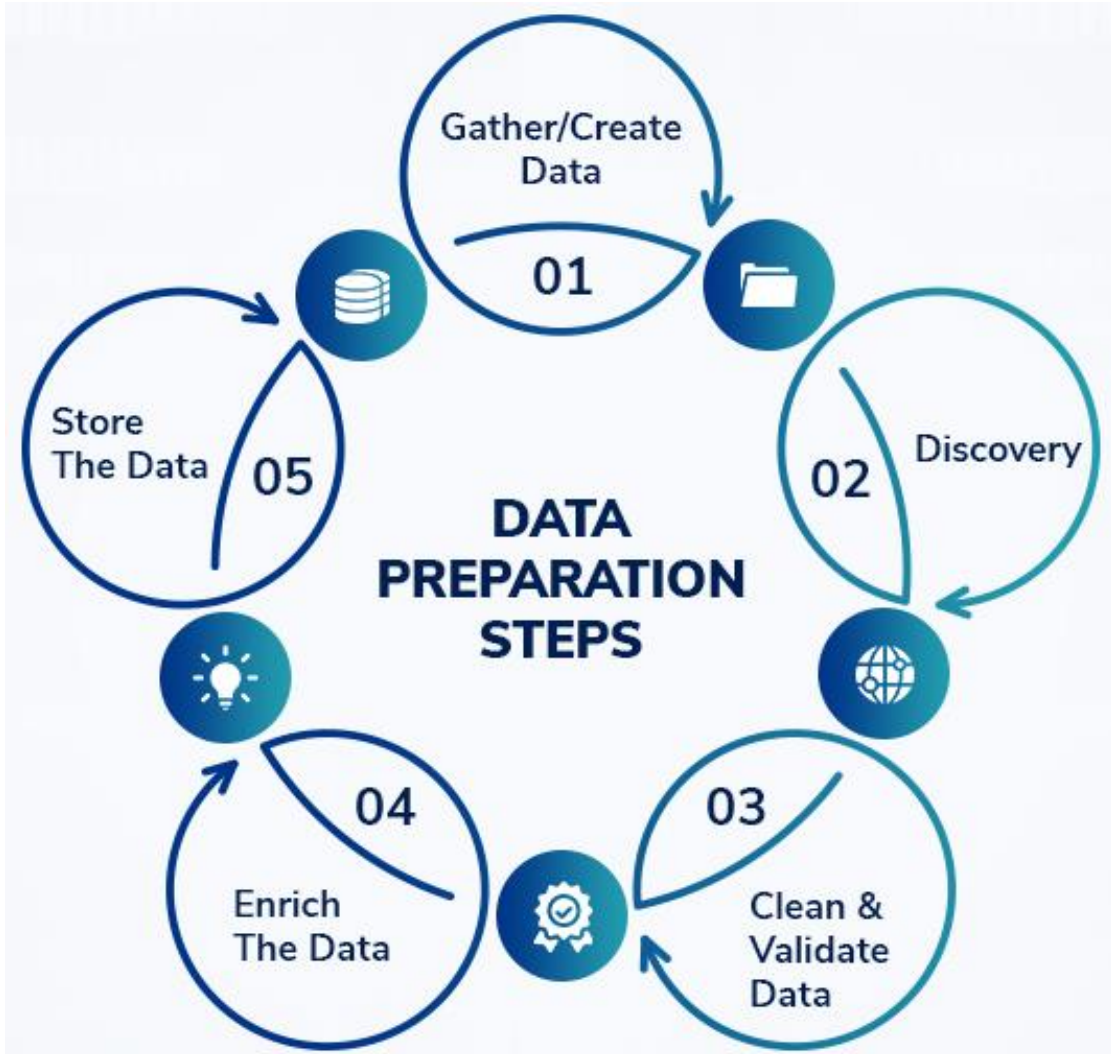
What is Data Engineering?



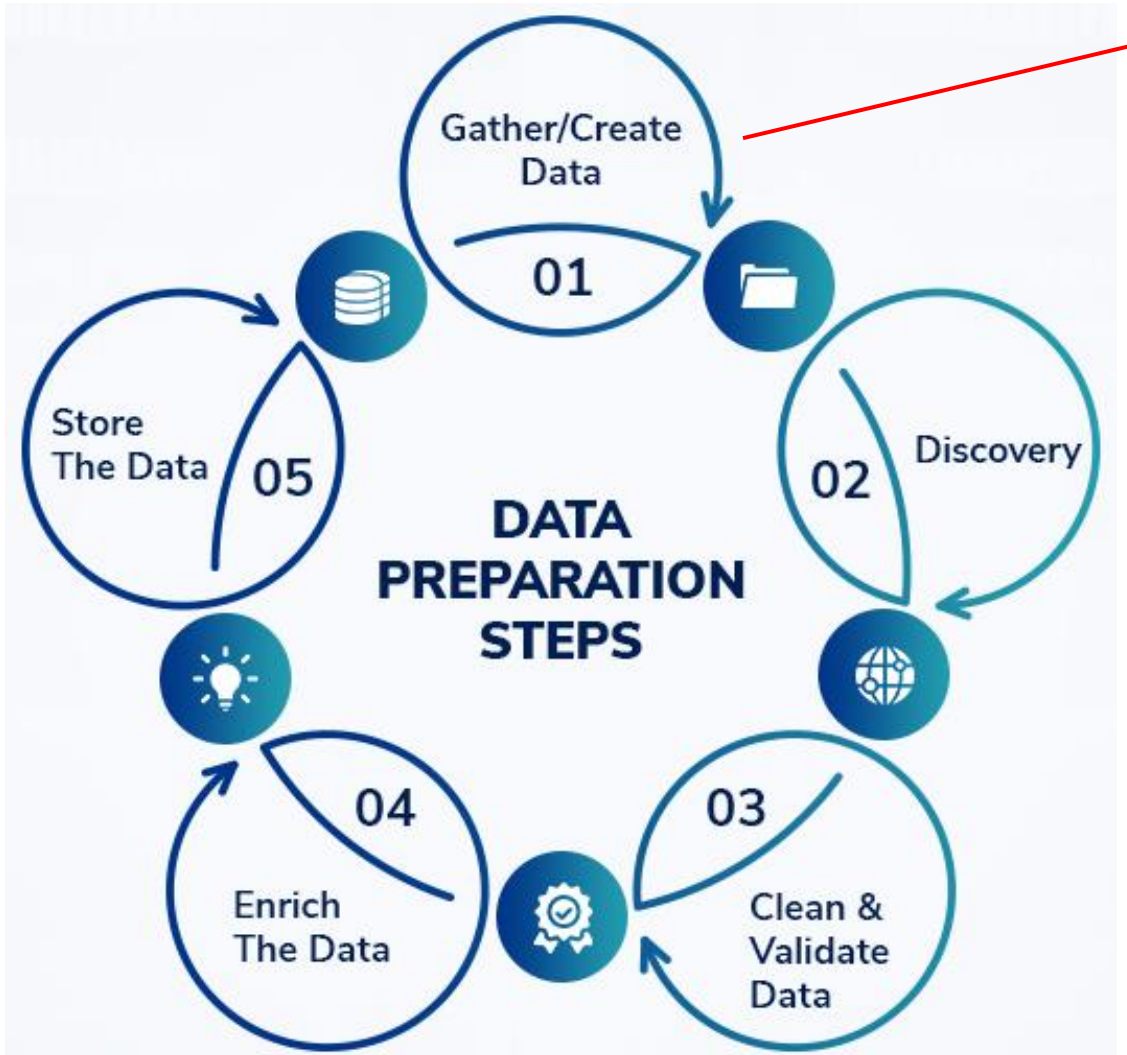
Course Overview



Course Overview

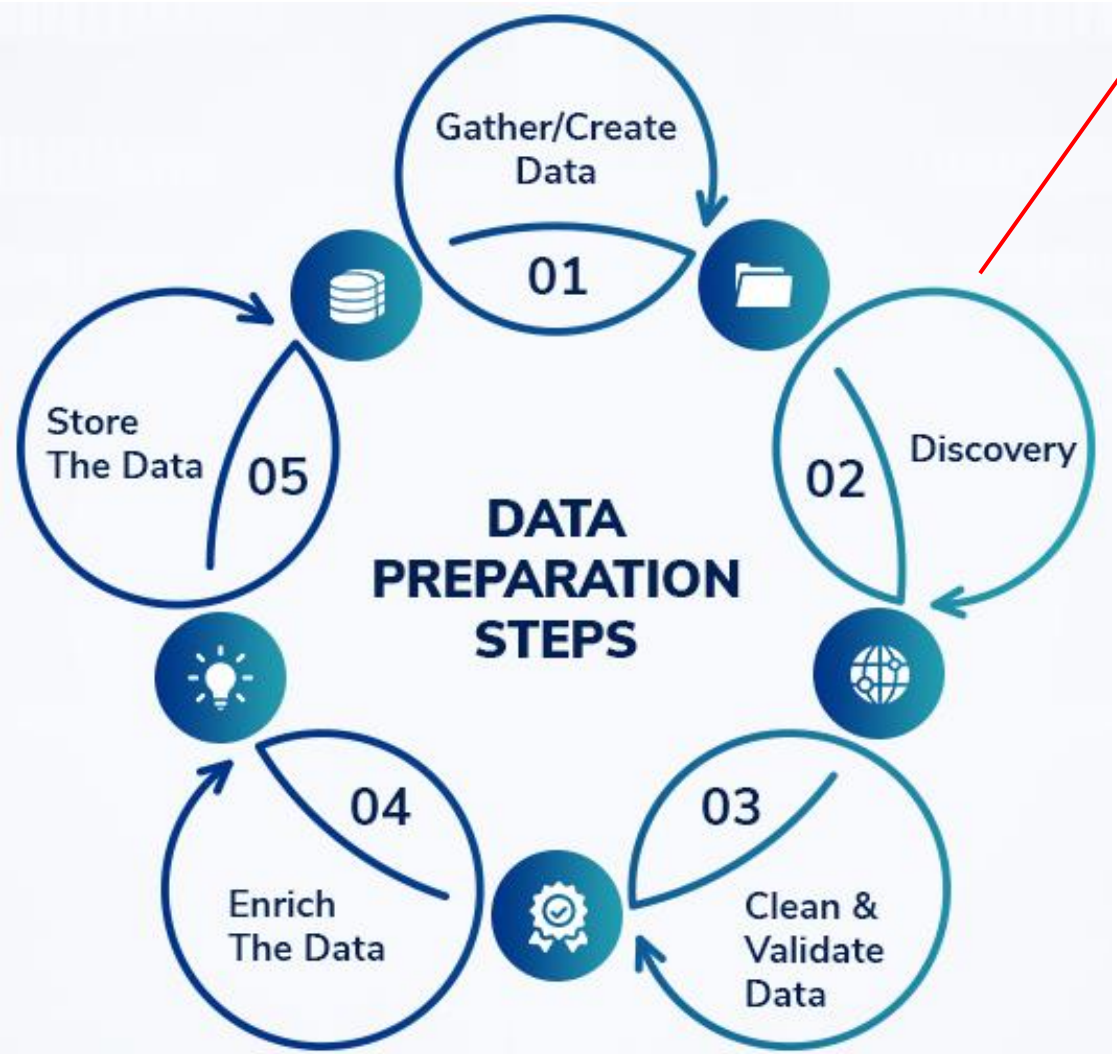


Course Overview



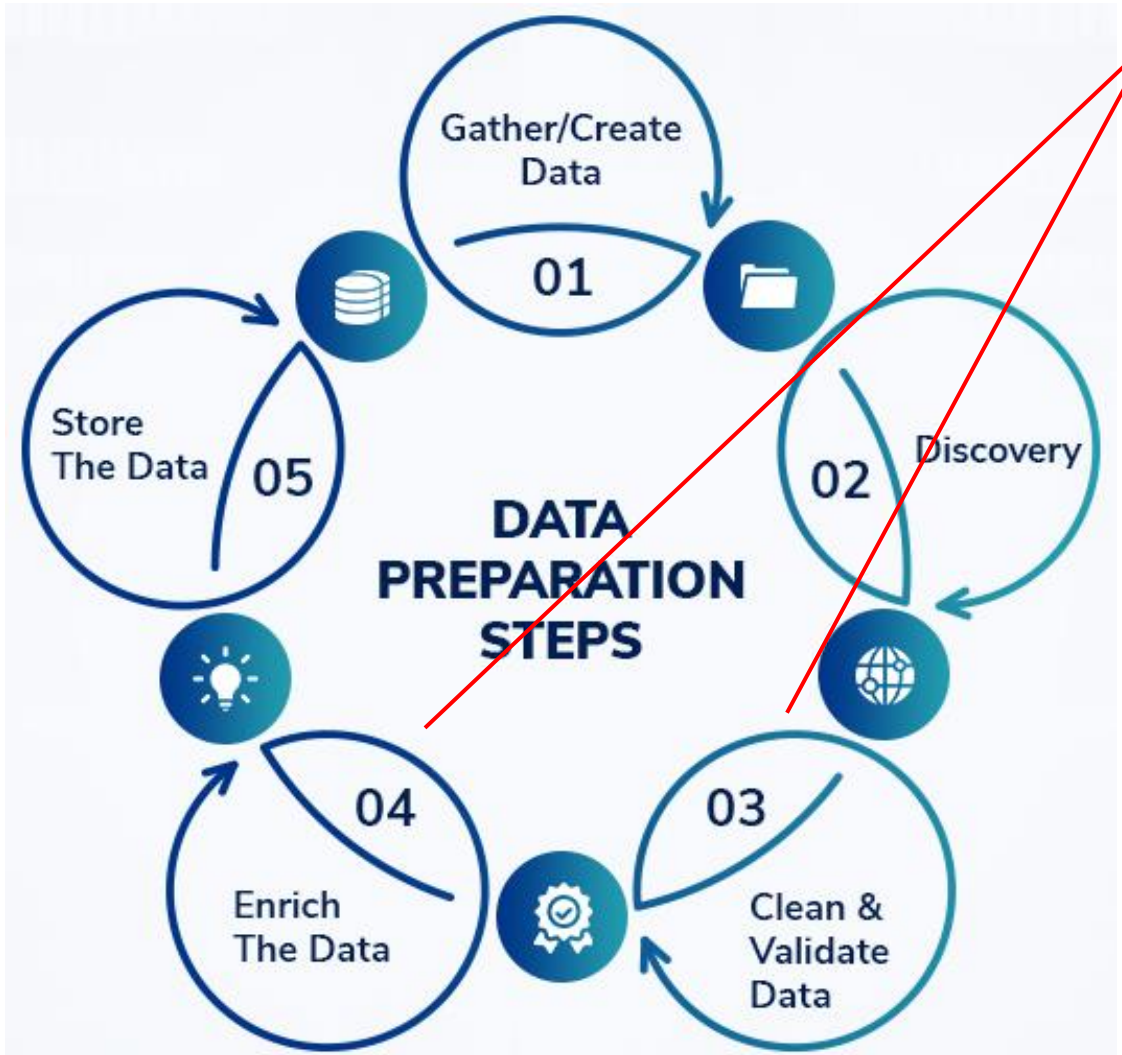
Lecture 2

Course Overview



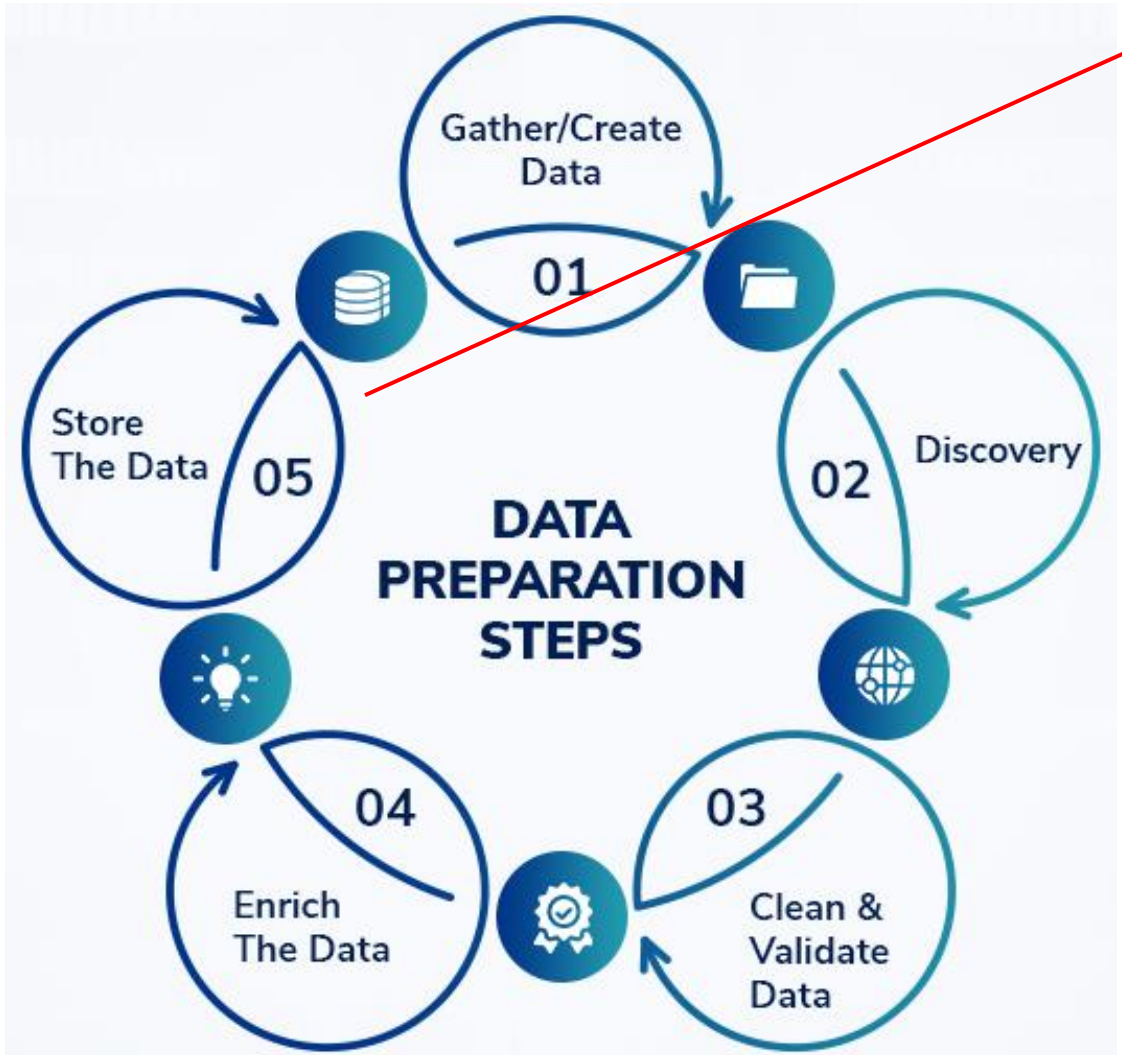
Lecture 3

Course Overview



Lecture 4-6

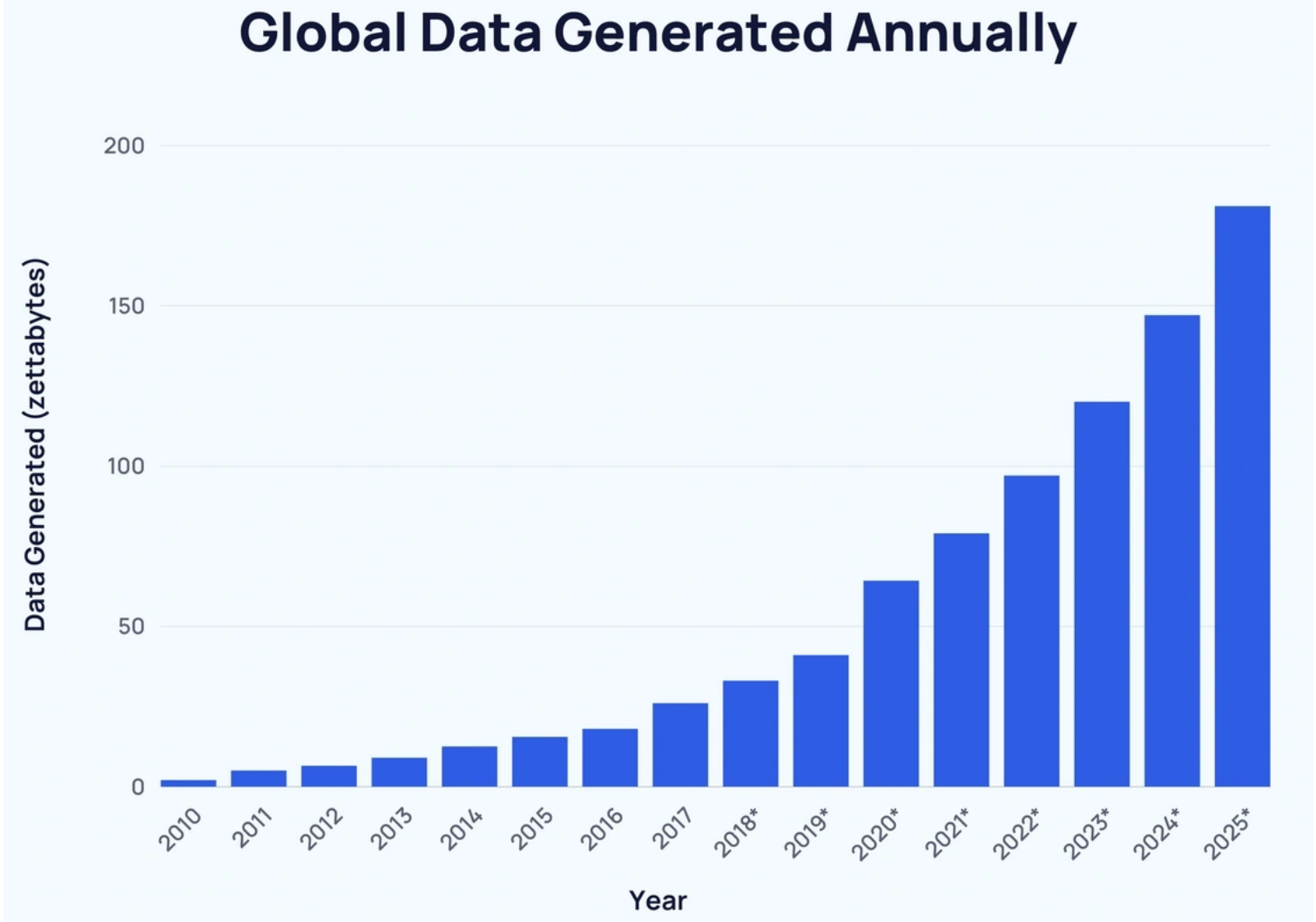
Course Overview



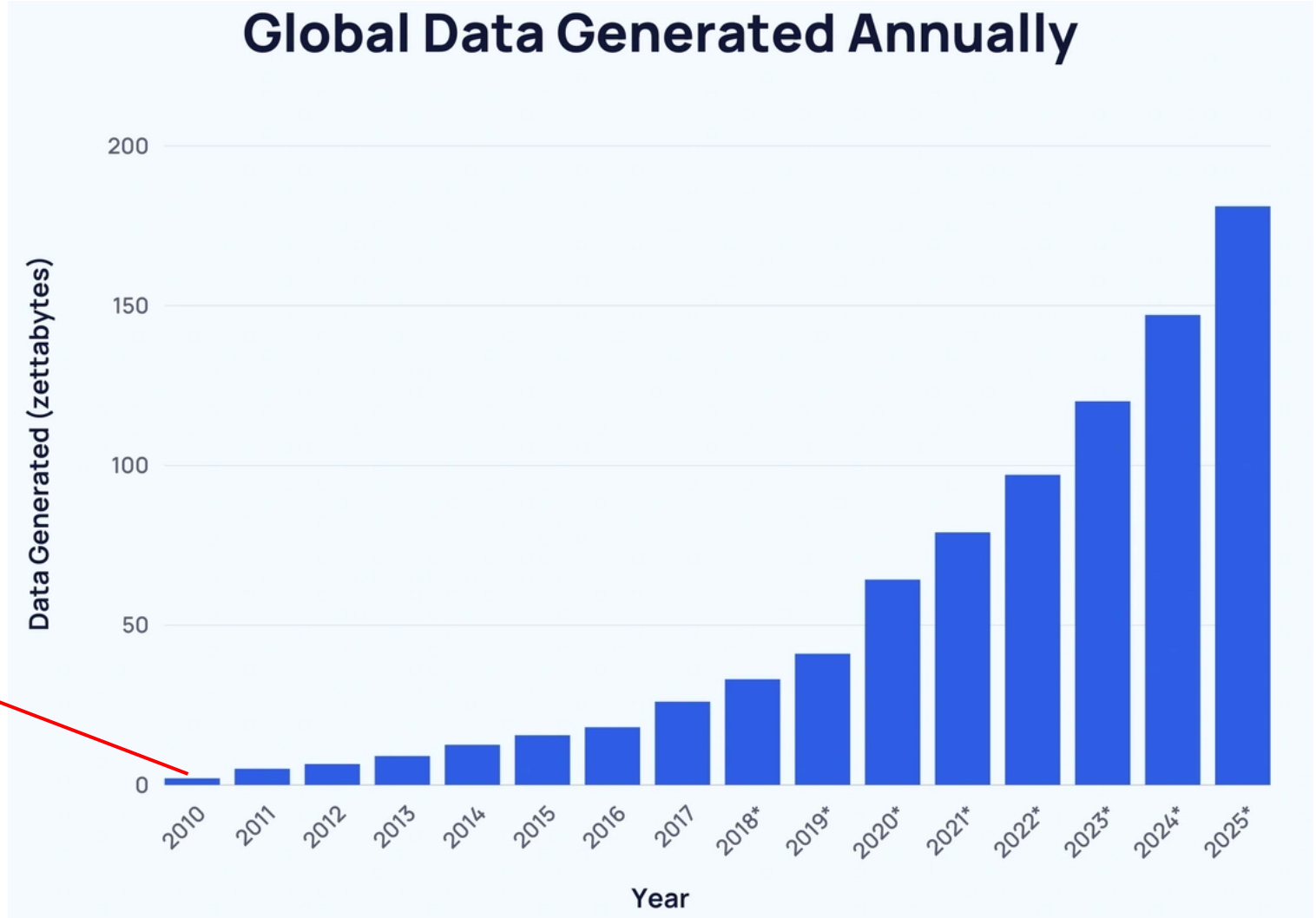
Lecture 7-15

Why do we need to study DB?

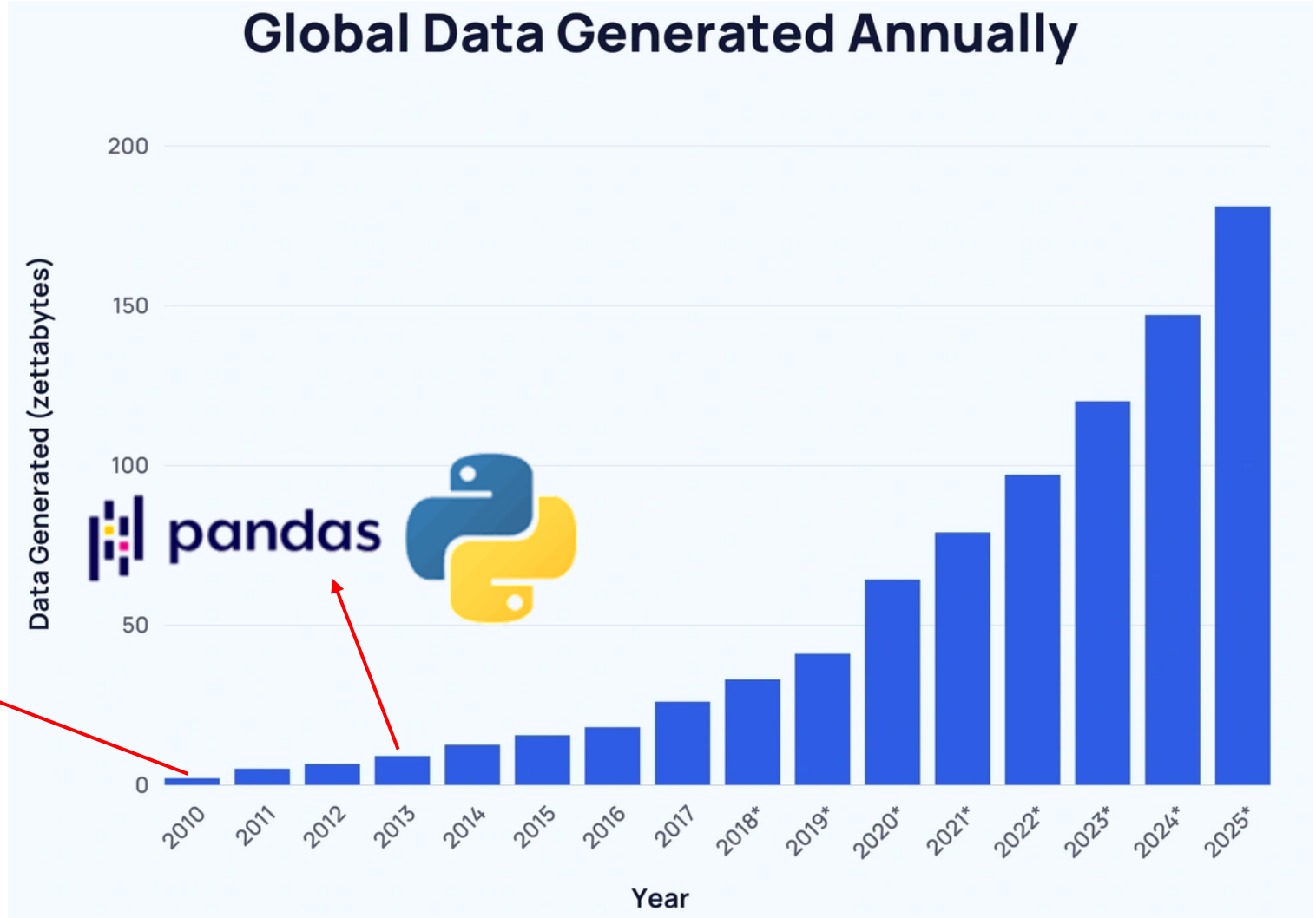
Why do we need to study DB?



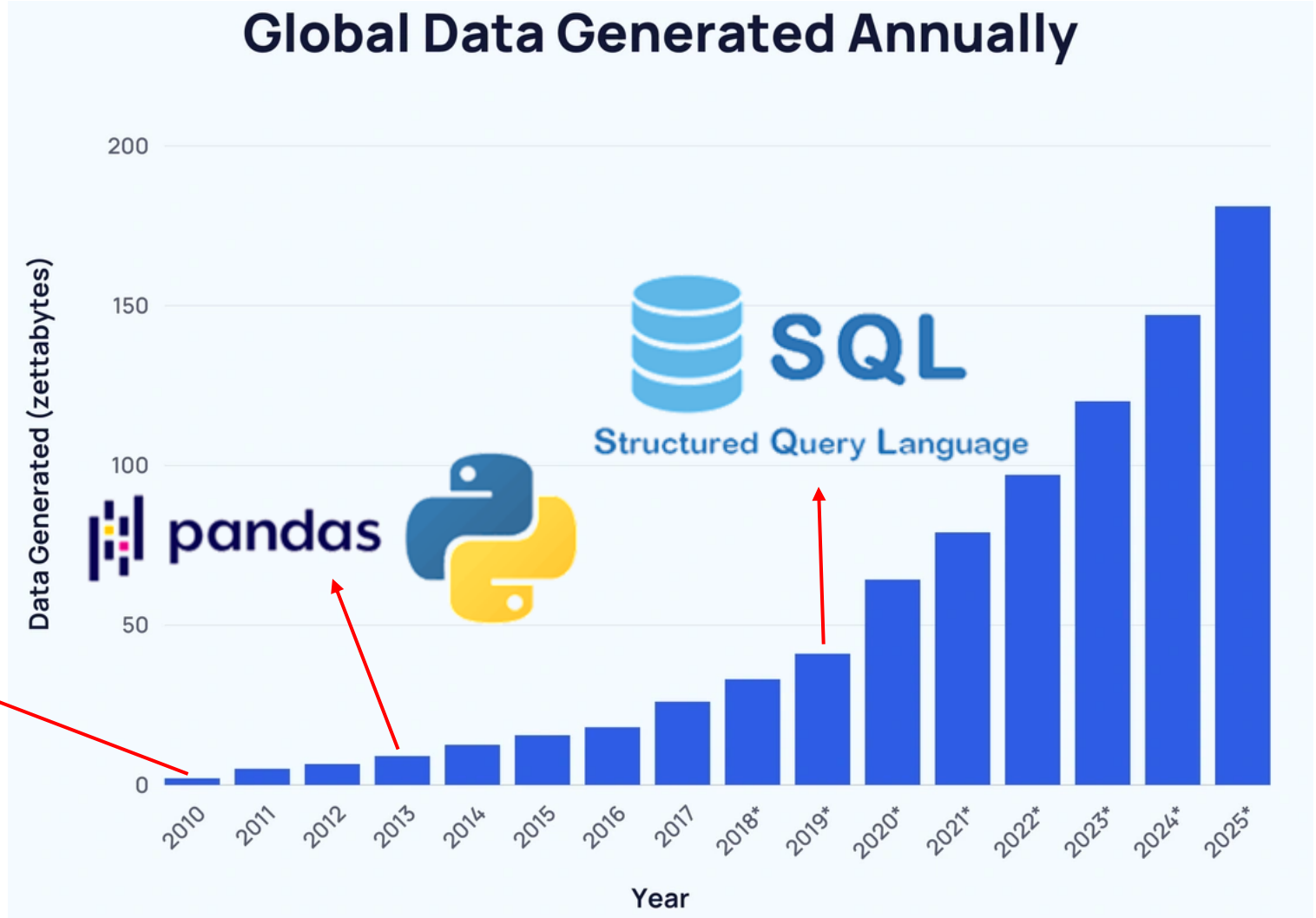
Why do we need to study DB?



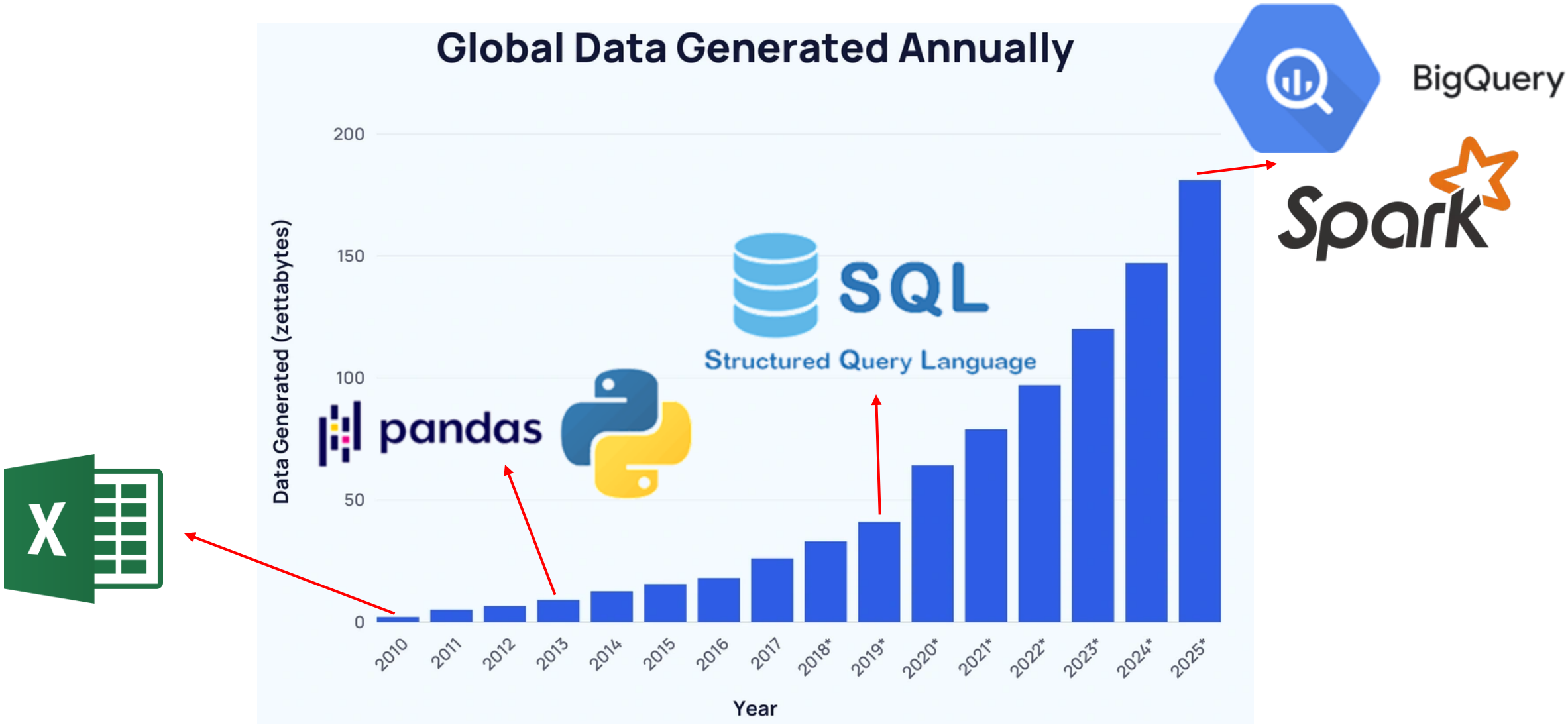
Why do we need to study DB?



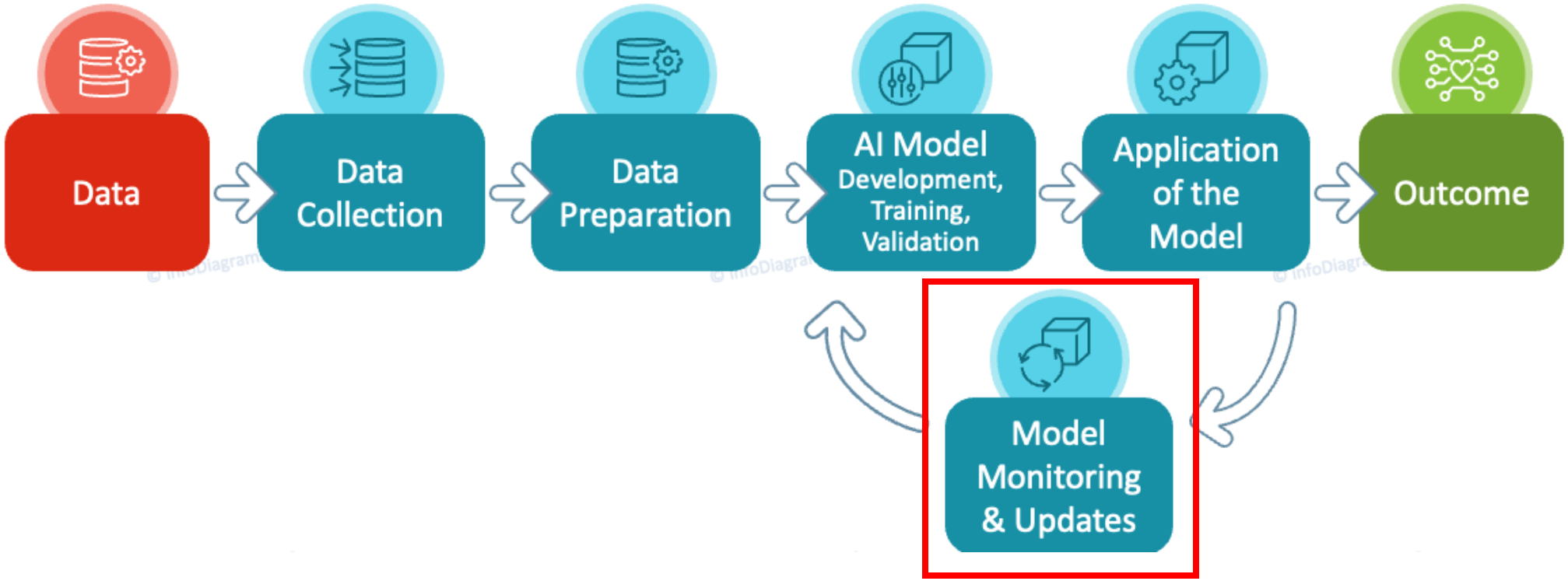
Why do we need to study DB?



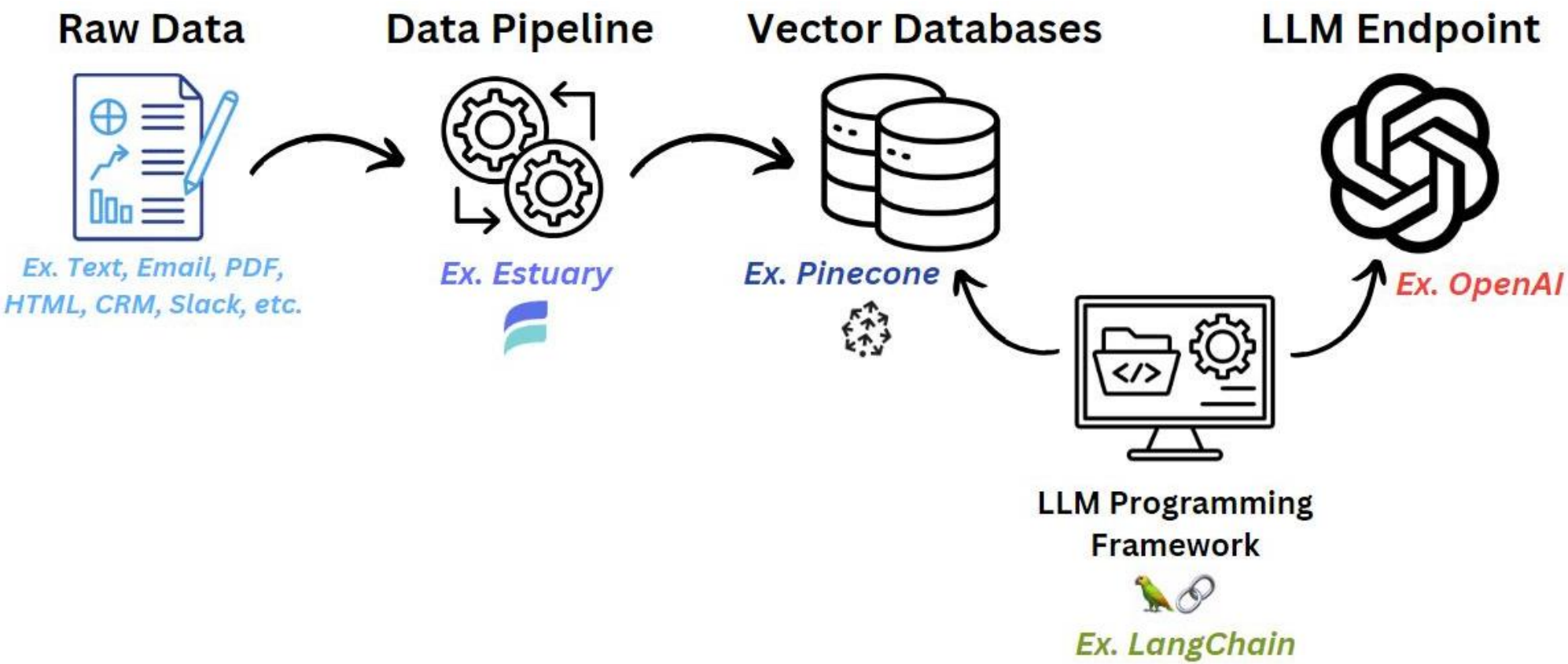
Why do we need to study DB?



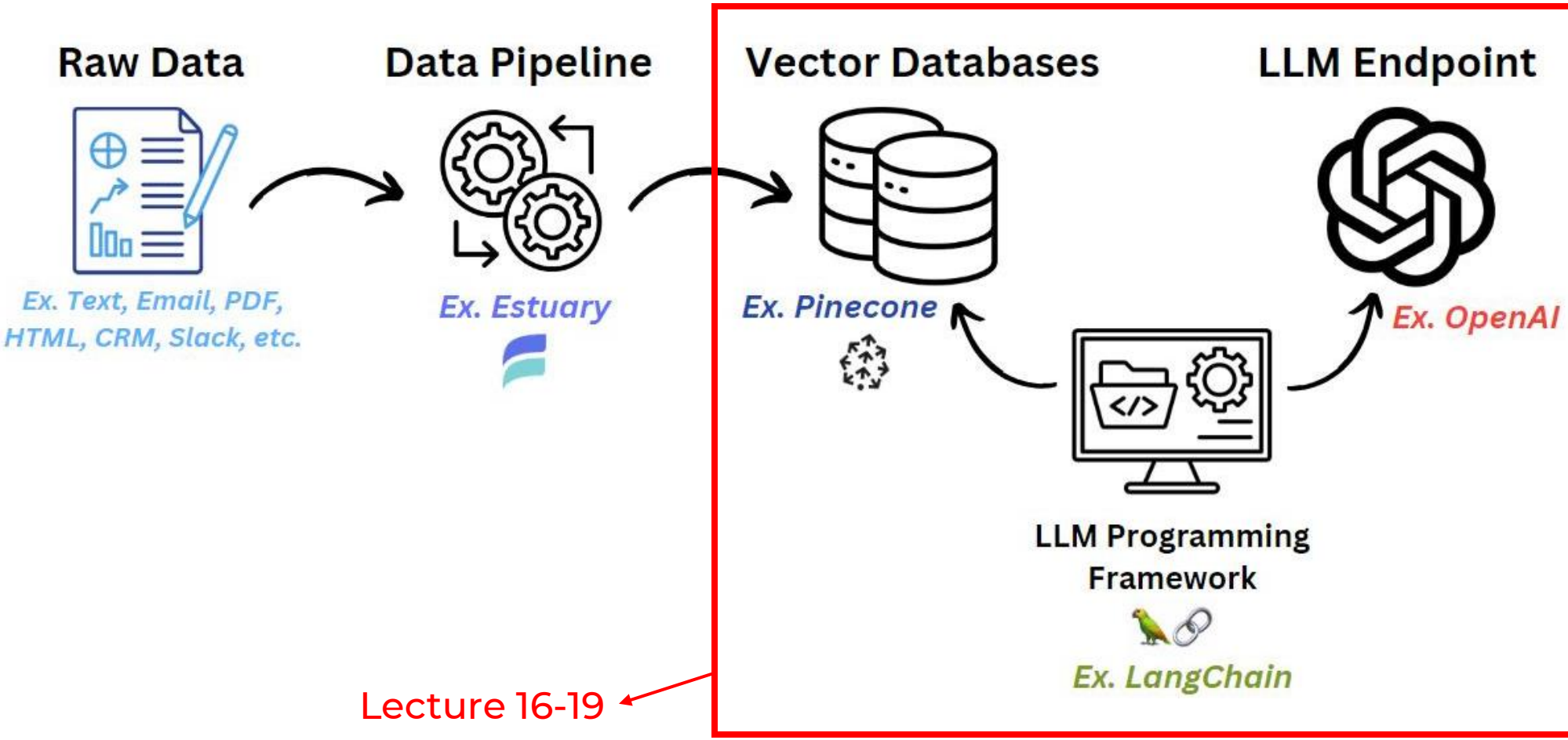
Course Overview (Cont'd)



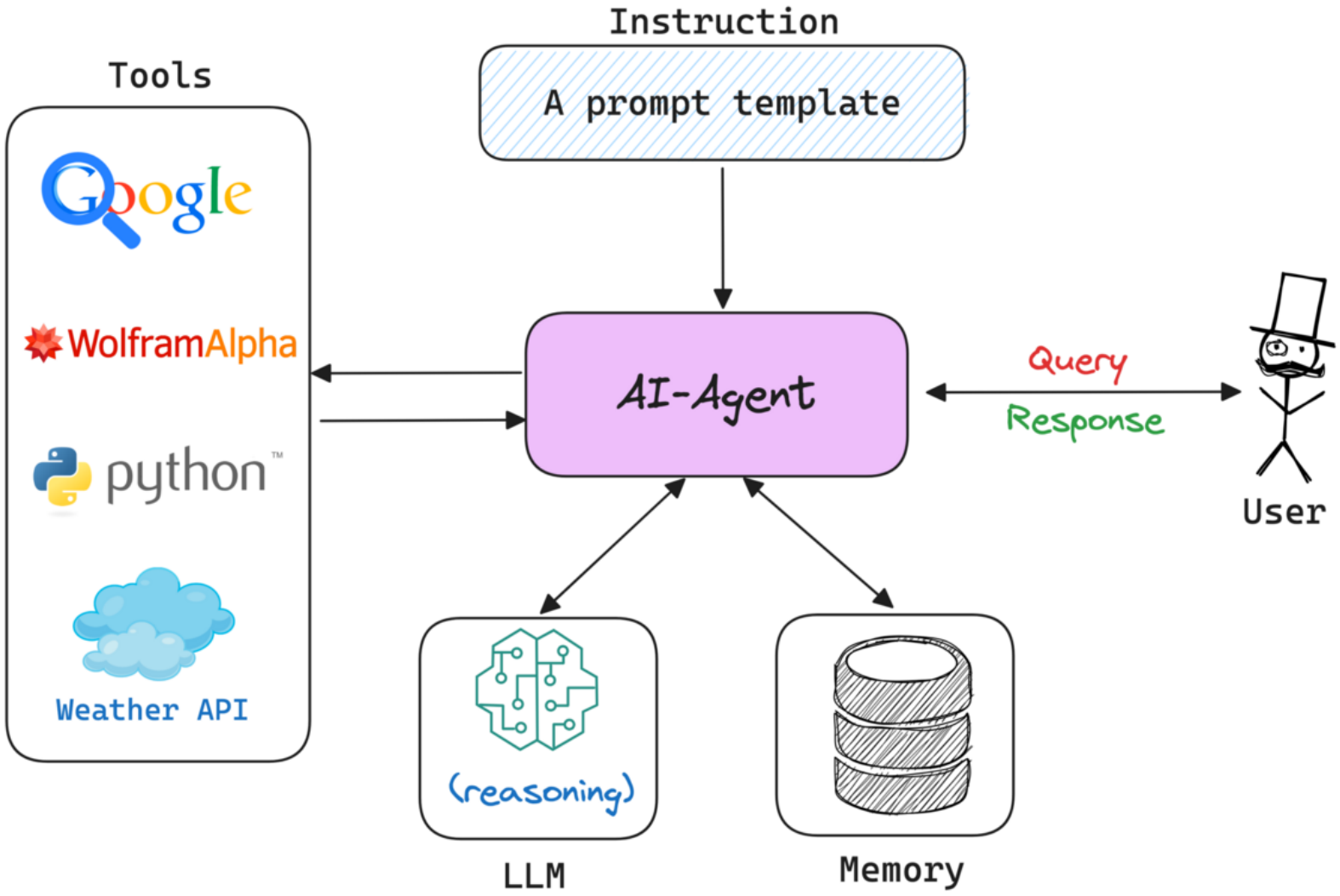
Course Overview (Cont'd)



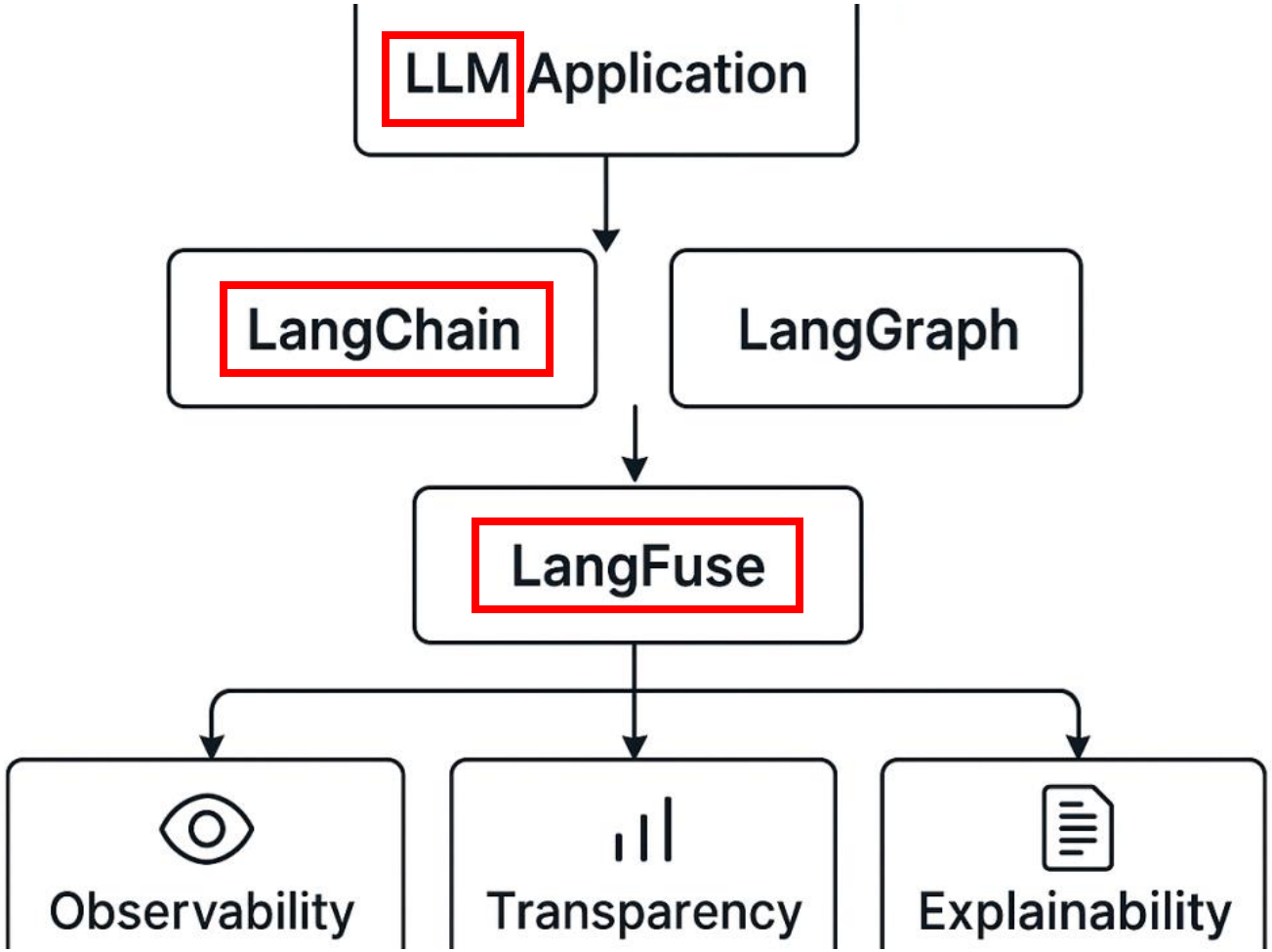
Course Overview (Cont'd)



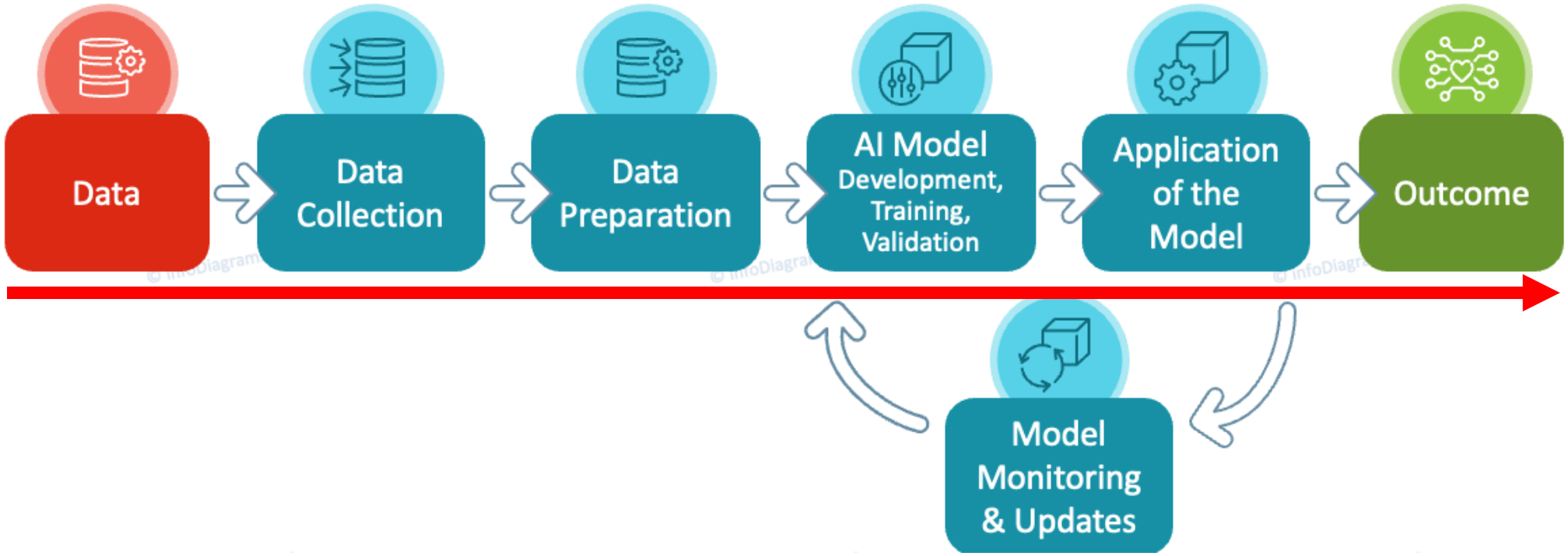
AI Agent + DB


Instruction
A prompt template
Tools
Google
WolframAlpha
python
Weather API
AI-Agent
Query
Response
User
(reasoning)
LLM
Memory

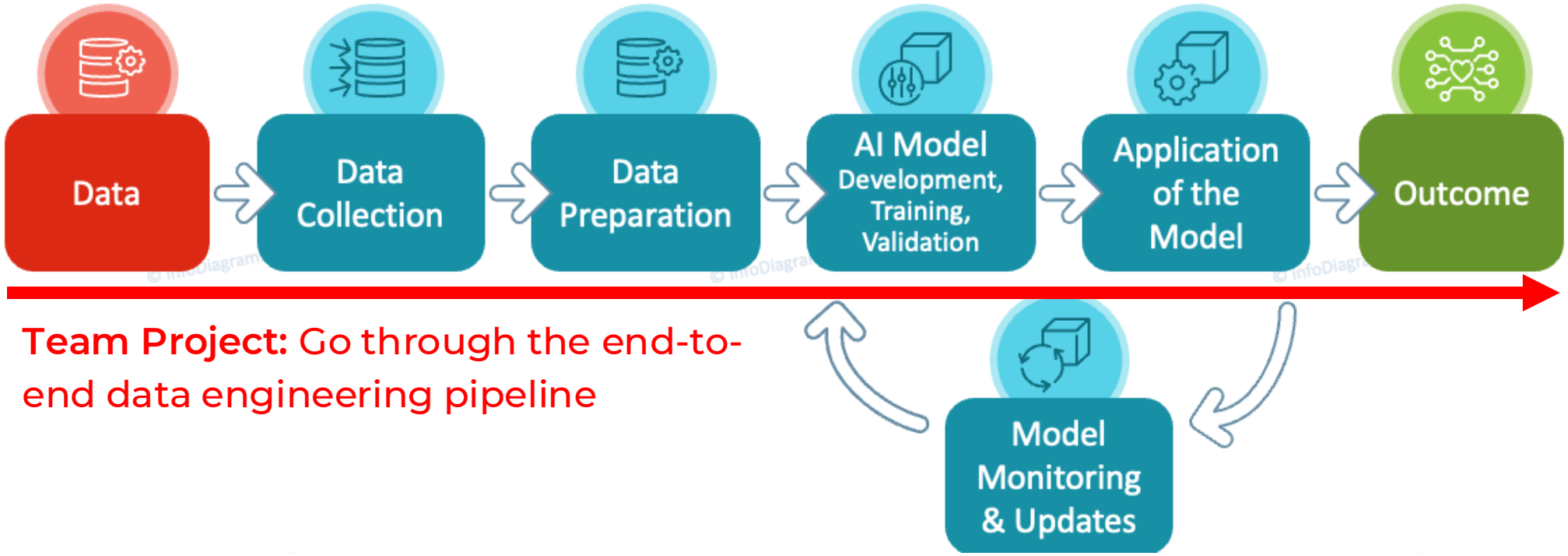
AI Agent + DB


LLM Application
LangChain
LangGraph
LangFuse
Observability
Transparency
Explainability

By the end of the class,



By the end of the class,



Team Project: Go through the end-to-end data engineering pipeline

Grading Policy

- Attendance (**10%**)
- Midterm Exam (**25%**)
- Final Exam (**25%**)
- Team Project (**40%**)

Office Hours: **Thursdays 11:00-12:00 pm, Room 319**

Team Project

Evaluation Criteria

- **Novelty:** originality and creativity of the idea
- **Impact:** relevance and potential usefulness of the solution
- **Technology:** effective use of the techniques and tools covered in class
- **Presentation:** clarity and professionalism of the report and presentation

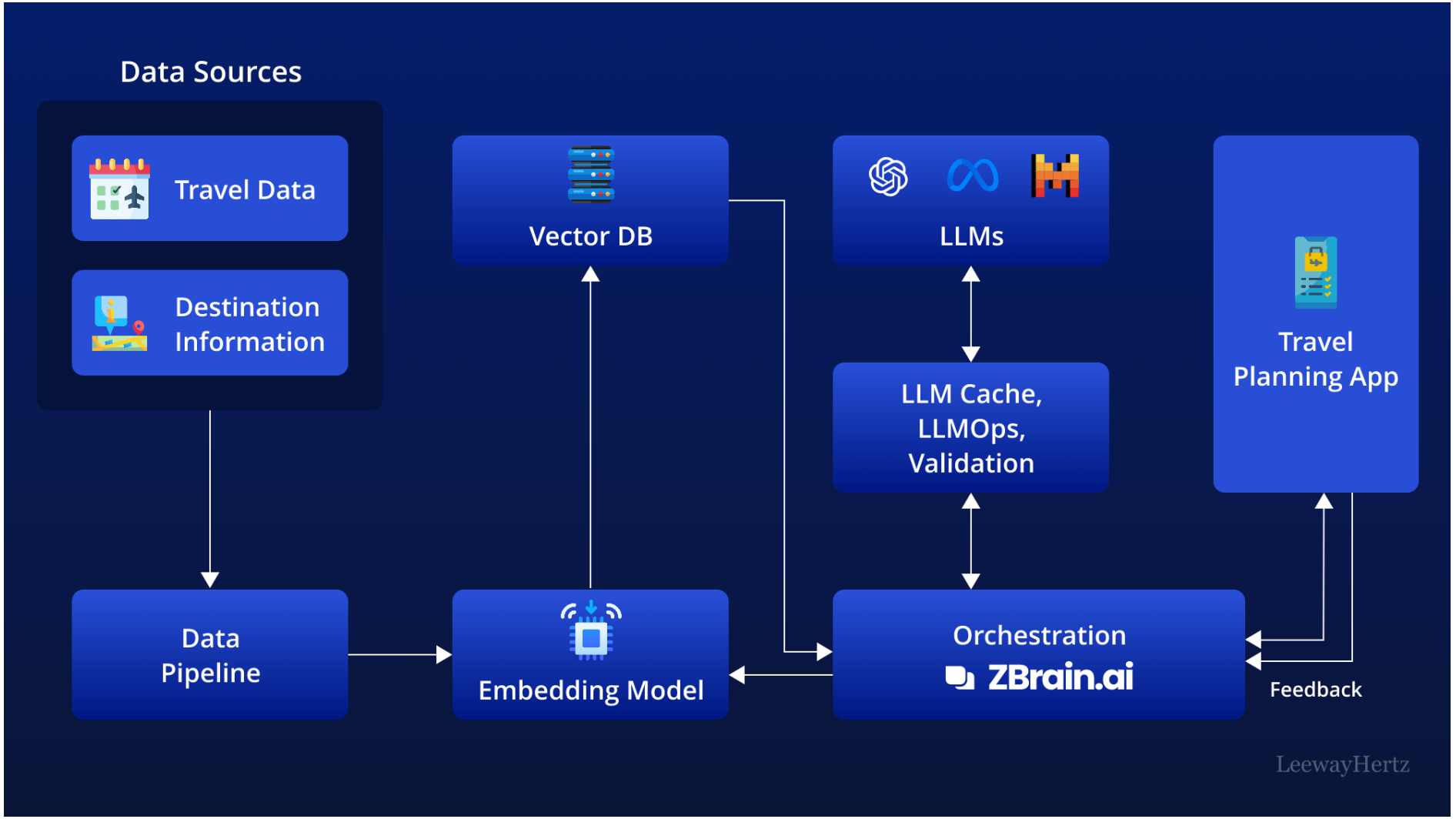
Team Size: **3~5**

Team Project

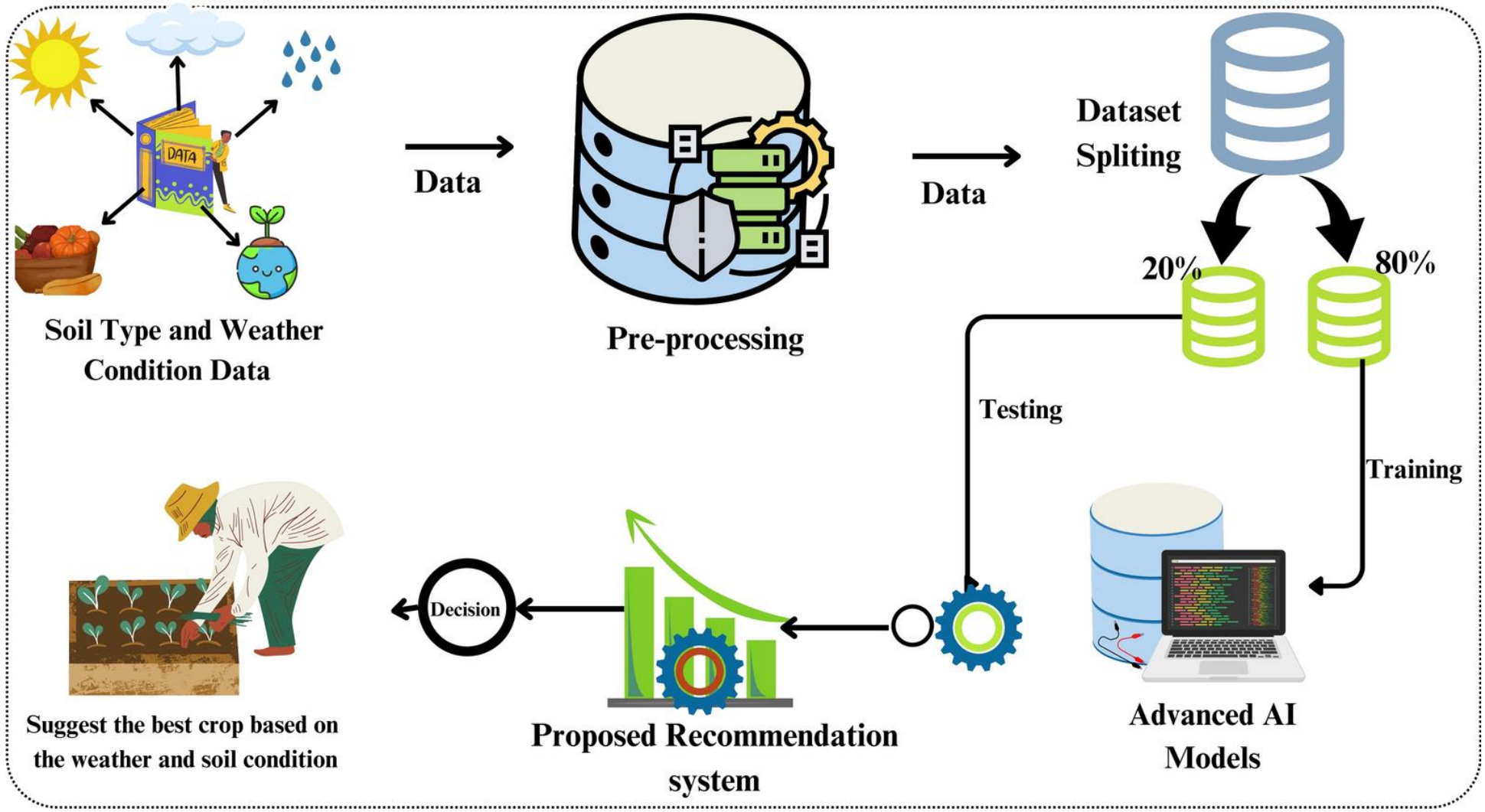
- **Proposal** (mid-semester):
  - Ideation, data collection & preprocessing, DB management

- **Final** (end-semester):
  - Develop a product that addresses a real-world problem using a large-scale DB (with potential LLM integration)

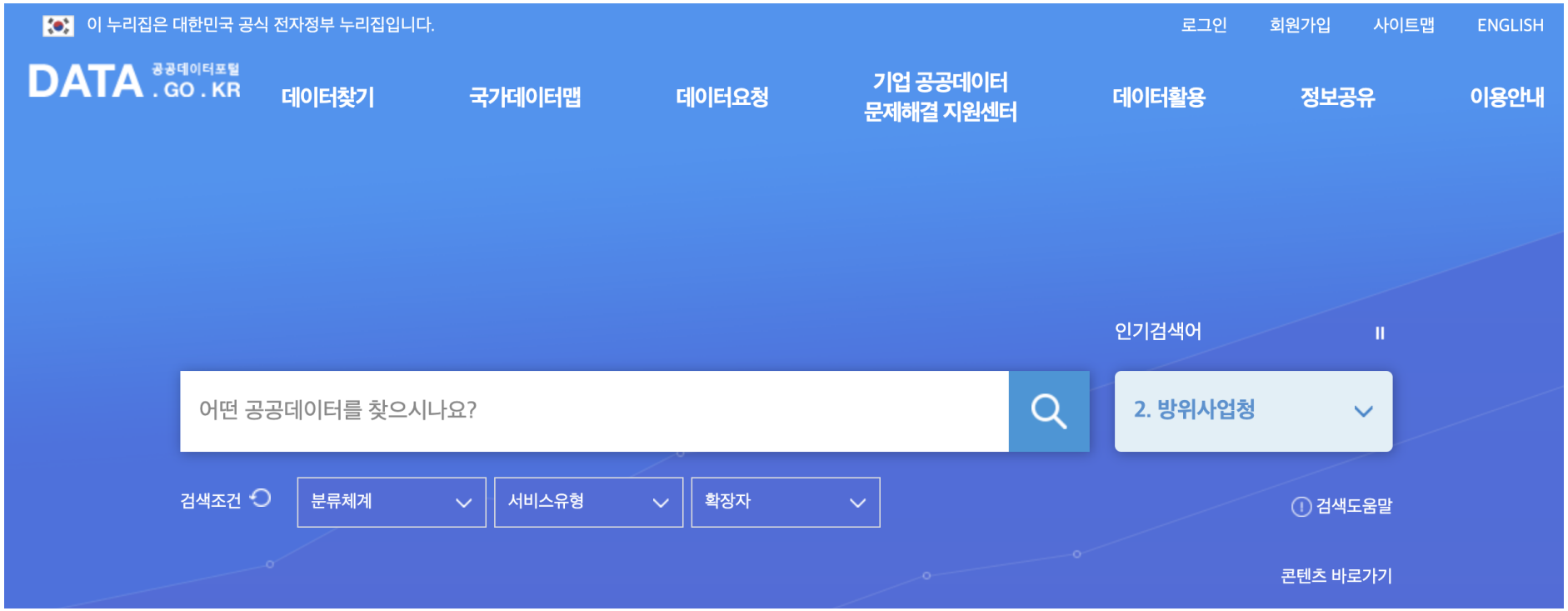
Team Project – Example Outcomes



Team Project – Example Outcomes



Team Project – Data Sources



AI Policy



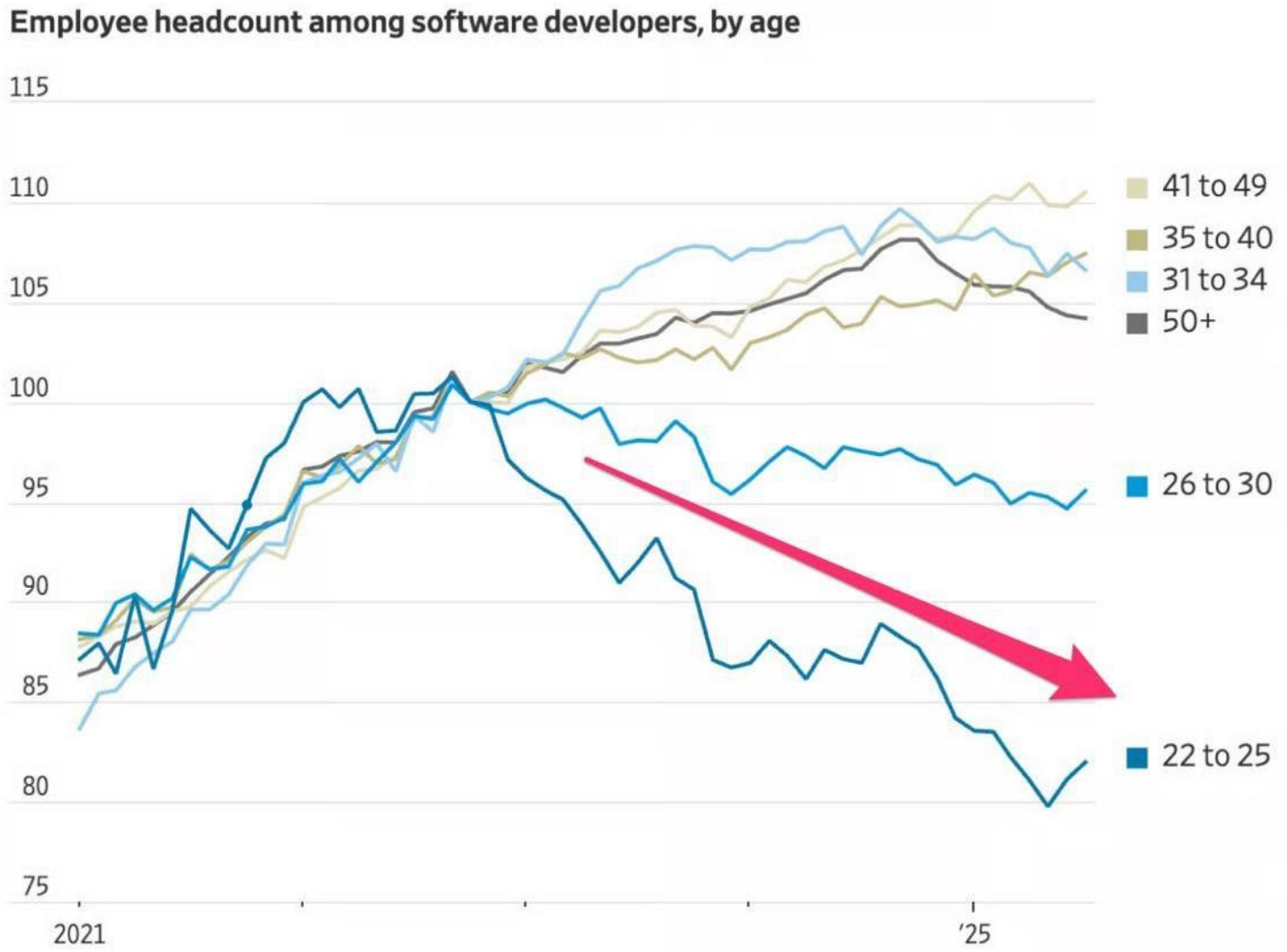
AI Policy



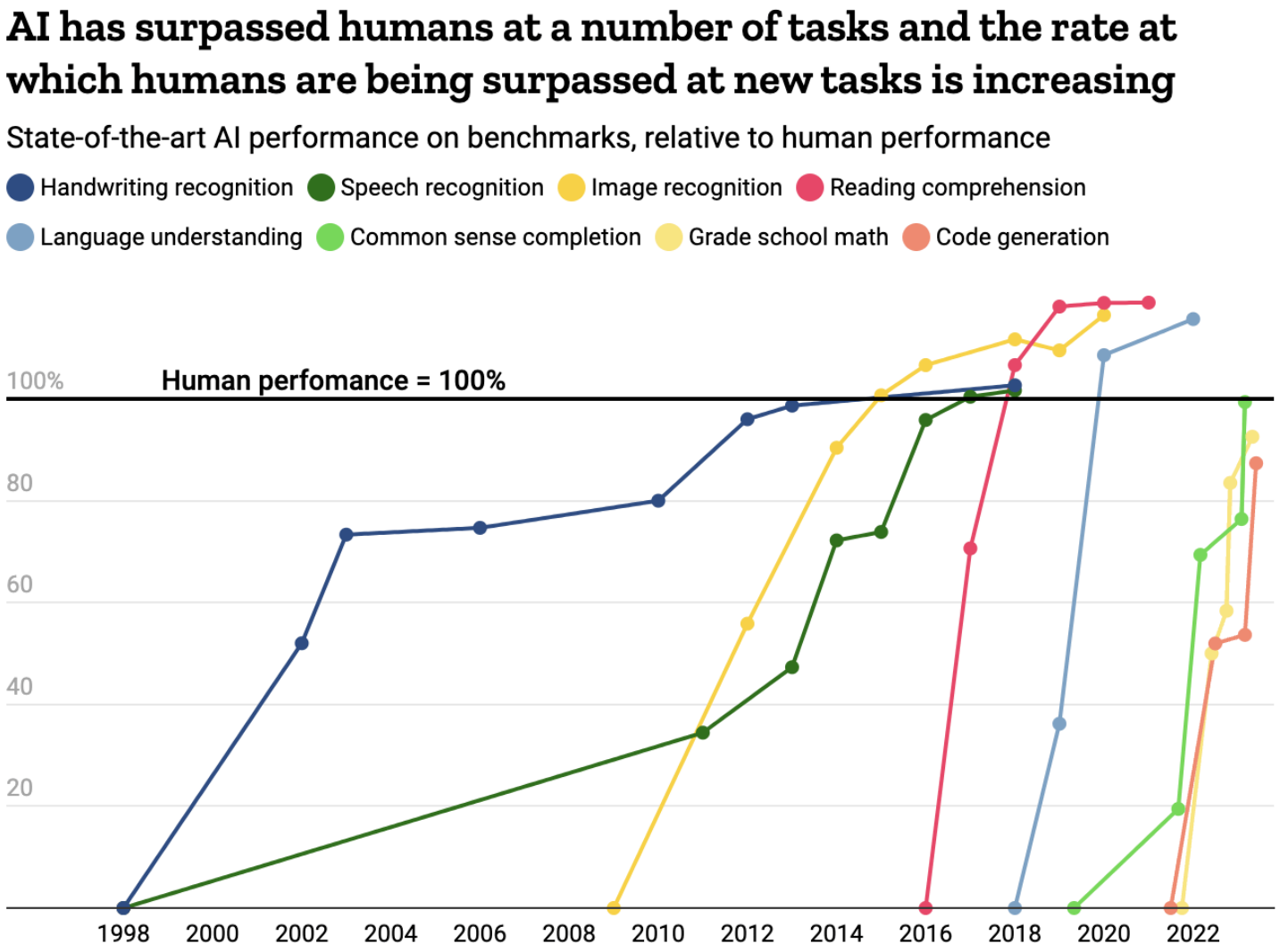
ChatGPT

Do not **compete** with AI, **utilize** it!

AI takes over entry-level jobs



Knowing is NOT enough



We must APPLY



Knowing is not enough;
we must apply.
Willing is not enough;
we must do

Johann Wolfgang von
Goethe

Schedule in the Google Sheets

#	#	Date		Topic	Slide	Reference/Note
	1	9/2	1	Overview & Logistics	https://drive.google.com/file/d/15	
	2	9/4	2	Data Collection	https://drive.google.com/file/d/1C	
	3	9/9	3	Data Exploration	https://drive.google.com/file/d/1S	DMCT Ch 2
	4	9/11	4	Data Pre-processing I: Cleaning, Integration	https://drive.google.com/file/d/1B	DMCT Ch 3
	5	9/16	5	Data Pre-processing II: Transformation	https://drive.google.com/file/d/1X	DMCT Ch 3
	6	9/18	6	Data Pre-processing III: Reduction	https://drive.google.com/file/d/1B	DMCT Ch 3
	7	9/23		Canceled		
	8	9/25	7	SQL I: Basics	https://drive.google.com/file/d/1V	DMS Lec 1-2
	9	9/30	8	SQL II: Joins, Multisets	https://drive.google.com/file/d/13	DMS Lec 2-3
	10	10/2	9	SQL III: Subqueries, Aggregations	https://drive.google.com/file/d/1M	DMS Lec 3-4
	11	10/7		Chuseok		
	12	10/9		Chuseok		
	13	10/14		Term Project Proposal I		
	14	10/16		Term Project Proposal II		
	15	10/21		Study day		
	16	10/23		Mid-term exam		